Benchmarking Robustness of AI-enabled Multi-sensor Fusion Systems: Challenges and Opportunities

Xinyu Gao
xinyugao@smail.nju.edu.cn
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

Zhijie Wang
zhijie.wang@ualberta.ca
University of Alberta
Edmonton, AB, Canada

Yang Feng[*]
fengyang@nju.edu.cn
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

Lei Ma[*]
ma.lei@acm.org
The University of Tokyo, Japan
University of Alberta, Canada

Zhenyu Chen
zychen@nju.edu.cn
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

Baowen Xu
bwxu@nju.edu.cn
State Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210023, China

ABSTRACT

Multi-Sensor Fusion (MSF) based perception systems have been the foundation in supporting many industrial applications and domains, such as self-driving cars, robotic arms, and unmanned aerial vehicles. Over the past few years, the fast progress in data-driven artificial intelligence (AI) has brought a fast-increasing trend to empower MSF systems by deep learning techniques to further improve performance, especially on intelligent systems and their perception systems. Although quite a few AI-enabled MSF perception systems and techniques have been proposed, up to the present, limited benchmarks that focus on MSF perception are publicly available. Given that many intelligent systems such as self-driving cars are operated in safety-critical contexts where perception systems play an important role, there comes an urgent need for a more in-depth understanding of the performance and reliability of these MSF systems.

To bridge this gap, we initiate an early step in this direction and construct a public benchmark of AI-enabled MSF-based perception systems including three commonly adopted tasks (i.e., object detection, object tracking, and depth completion). Based on this, to comprehensively understand MSF systems' robustness and reliability, we design 14 common and realistic corruption patterns to synthesize large-scale corrupted datasets. We further perform a systematic evaluation of these systems through our large-scale evaluation and identify the following key findings: (1) existing AI-enabled MSF systems are not robust enough against corrupted sensor signals; (2) small synchronization and calibration errors can lead to a crash of AI-enabled MSF systems; (3) existing AI-enabled MSF systems are usually tightly-coupled in which bugs/errors from an individual sensor could result in a system crash; (4) the robustness of MSF systems can be enhanced by improving fusion mechanisms. Our results reveal the vulnerability of the current AI-enabled MSF perception systems, calling for researchers and practitioners to take robustness and reliability into account when designing AI-enabled MSF. Our benchmark, code, and detailed evaluation results are publicly available at https://sites.google.com/view/ai-msf-benchmark.

CCS CONCEPTS

• **Software and its engineering → Software defect analysis**; • **General and reference → Empirical studies**.

KEYWORDS

Multi-Sensor Fusion, Benchmarks, AI Systems, Perception Systems

ACM Reference Format:
Xinyu Gao, Zhijie Wang, Yang Feng, Lei Ma, Zhenyu Chen, and Baowen Xu. 2023. Benchmarking Robustness of AI-enabled Multi-sensor Fusion Systems: Challenges and Opportunities. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23), December 3–9, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3611643.3616278

1 INTRODUCTION

Multi-sensor fusion (MSF) refers to the technique that combines data from multiple sources of sensors to achieve specific tasks, which has been widely adopted in many real-world complex systems. The integration of information from different sensors avoids the inherent perception limitations of individual sensors and improves the system's overall performance. Over the past years, MSF-based perception systems have been widely used in various industrial domains and safety-critical applications, such as self-driving cars [24], unmanned aerial vehicles [36], and robotic systems [29].

With recent advances in data-driven artificial intelligence (AI), there comes an increasing trend in proposing deep learning (DL) techniques to further enable more advanced heterogeneous data

[*] Yang Feng and Lei Ma are the corresponding authors.

ESEC/FSE '23, December 3–9, 2023, San Francisco, CA, USA

ACM ISBN 979-8-4007-0327-0/23/12…$15.00
https://doi.org/10.1145/3611643.3616278

processing from different sensors, in order to achieve more accurate perception and prediction. Given the advantage of deep neural networks (DNNs) in processing and extracting complex semantic information from sensors' data (e.g., image, point cloud), AI-enabled MSF has been increasingly adopted in the perception systems of autonomous driving [24, 36].

The rapid development of AI-enabled MSF systems also brings challenges and concerns. One of the biggest concerns is the lack of a deep understanding of the current AI-enabled MSF's reliability. In practice, an AI-enabled MSF system could behave incorrectly and lead to severe accidents in safety-critical contexts, especially in autonomous driving [10, 51]. Thus, it is highly desirable to enable testing, analysis, and systematic assessment of such intelligent systems beforehand comprehensively. One common practice to enable such quality assurance activities in AI/SE communities is to establish a benchmark that enables both researchers and practitioners to perform systematic studies and develop novel techniques to better fulfill important quality requirements. However, to the best of our knowledge, up to the present, few benchmarks specifically designed for AI-enabled MSF are yet available. It is unclear whether and to what extent the potential quality issues and risks can be, how they are brought from each sensing unit, and their impacts on the integration and the state-of-the-art information fusion processes.

To bridge this gap, in this paper, we initiate an early step to present a benchmark and perform an empirical study of AI-enabled MSF perception systems. Fig. 1 summarizes the high-level design and workflow of our benchmark construction and our empirical study, in which we mainly investigate the following research questions, aiming to identify the potential challenges and opportunities:

- **RQ1. How do AI-enabled MSF-based perception systems perform against common corrupted signals?** This RQ aims to investigate the potential risks of AI-enabled MSF systems against corrupted signals that commonly occur in the operational environments. Through a large-scale evaluation on eleven types of corrupted sensor signals, we find that the current AI-enabled MSF systems are not robust enough, especially against weather condition changes.
- **RQ2. How sensitive is AI-enabled MSF when facing spatial and temporal misalignment of sensors?** In the practical open and wild environment, it is almost impossible to always maintain perfect calibration or precise time synchronization of the system across sensors. RQ2 aims to investigate the sensitivity of AI-enabled MSF to spatial and temporal misalignment. Our experiment results reveal that even small calibration or synchronization issues could lead to abnormal behaviors of the system.
- **RQ3. To what extent are existing sensing components coupled of an AI-enabled MSF system?** A robust and reliable MSF should not completely fail when one or a part of the whole sensing modules lose the source signal. RQ3 aims to investigate how AI-enabled MSF systems can be impacted when one source of the signal is partially/completely lost. Overall, we find that the tightly-coupled architecture of AI-enabled MSF systems exhibits less robustness against signal loss.
- **RQ4. What is the weakness of different AI-enabled MSF mechanisms and is it possible to repair them?** RQ4 aims to investigate the unique advantages of each fusion mechanism and potential opportunities for improving the robustness of AI-enabled MSF systems. Our results demonstrate that deep fusion is more robust in some cases, however, weak and late fusion can be easier to be repaired in terms of robustness against corruption patterns.

To sum up, this work makes the following contributions:

- **Benchmark**. We initiate to create an early public benchmark of AI-enabled MSF-based perception systems. This provides a common ground for the study and analysis of AI-enabled MSF systems' robustness and enables future quality assurance research in this direction.
- **Empirical Study**. Based on the benchmark, we perform a large-scale empirical study of AI-enabled MSF systems to investigate their current status regarding robustness.
- **Discussion**. We further make discussions about existing AI-enabled MSF systems and future directions, including the unique advantages of different fusion mechanisms as well as the opportunities of their robustness enhancement.

To the best of our knowledge, this paper is among the very early research to benchmark and investigate the MSF system, which is a common and representative AI system composed of multiple sensing channels and corresponding models. On one hand, at present, it is not clear how much and to what extent each sensing unit could impact the integrated sensing results of an MSF; it is not clear how the issues of different sensing units and channels are involved and propagate to the final results of different MSF designs either. Creating a benchmark at the current stage enables to investigate these important questions quantitatively, which also enables further relevant quality assurance research along this direction. On the other hand, in general, MSF-based perception systems play a key role to enable autonomous and intelligent systems, which potentially has a big impact on many applications and domains. With the recent fast pace in transforming into the data-driven intelligent era, we believe an early stage benchmark and investigation of the current MSF systems empowered by deep learning would also benefit the practitioners in understanding the limitation and proposing better MSF engineering techniques, paving the path towards designing safe and reliable autonomous intelligent systems.

2 BACKGROUND

2.1 Perception Systems in Intelligent Systems

An intelligent system (e.g., a self-driving car, a robotic, an unmanned aerial vehicle) is usually a complex system composed of various subsystems. These subsystems are cooperated to ensure safe and reliable operations of the intelligent system. The perception system is one of the key components in an intelligent system, which is in charge of sensing and processing environmental information through sensors to perform crucial tasks, e.g., object detection and object tracking. The prediction results from perception systems are then propagated to other components in the intelligent system such as planning and control systems. The perception systems lay the foundation of the intelligent system's workflow in perceiving and understanding the environment, which also significantly impact the quality and reliability of the whole system.

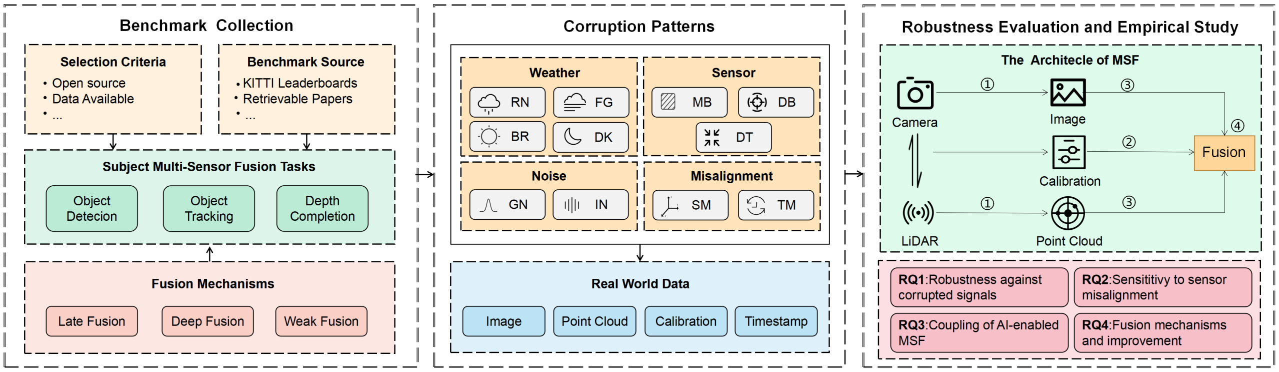


Figure 1: Workflow summary of AI-enabled MSF benchmark construction, and high-level empirical study design.

Most industrial-level systems leverage multi-sensor fusion (MSF) strategy to avoid inherent perception limitations of individual sensors and thus sense the environment more reliably [34]. For instance, the camera and LiDAR are usually fused in self-driving cars since camera is more effective in capturing semantic information and LiDAR could provide more accurate geographic information [9]. As shown in Fig. 1 (right part), each sensor in a camera-LiDAR fusion first senses the surrounding environment individually. Then, signals from different sensors are transformed into the same coordinate system and matched across the timestamps based on the temporal and spatial calibration among sensors. Finally, the fusion module receives calibrated and synchronized signals from different sensors, and fuses them to make predictions for downstream tasks.

2.2 AI-enabled Multi-Sensor Fusion

Different from traditional MSF that only fuses the data or output, AI-enabled MSF also has the possibility to fuse the deep semantic features learned by DNNs. We take the fusion of camera and LiDAR as an example (right part of Fig. 1) in the following sections.

Each branch that processes signals in AI-enabled MSF can be represented as a composite function chain (Eq. 1) that maps the input data M to the output result f^L.

$$f^L = F^{(L)}(F^{(L-1)}(\cdots(F^{(0)}(M)))), \tag{1}$$

where L denotes the depth of a branch. The medium output in the chain, i.e., $F^{(j)}(\cdot), j = \{1, 2, \ldots, L-1\}$, represents the output from jth hidden layer in a DNN.

Based on the stage where the fusion is made [12, 20] (see Fig. 2), AI-enabled MSF can be categorized into four different mechanisms at a high level: *early fusion*, *late fusion*, *deep fusion*, and *weak fusion*. Since *early fusion* is not commonly used in AI-enabled MSF, we focus on the other three fusion mechanisms in the rest of this paper. For a L layer deep neural network, we denote M_i and M_j as two different modalities and define $\oplus$ as a fusion operation. Now we briefly introduce each MSF mechanism.

Late fusion directly combines the output results of each branch, which can be formulated as:

$$\begin{aligned} f^L = F^{(L)}(F_i^{(L-1)}(\cdots(F_i^{(0)}(M_i)))) \\ \oplus F^{(L)}(F_j^{(L-1)}(\cdots(F_j^{(0)}(M_j)))) \end{aligned} \tag{2}$$

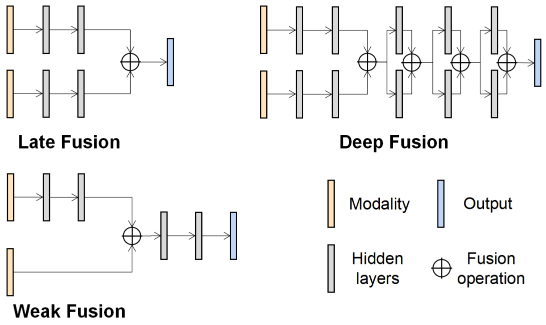


Figure 2: Different AI-enabled MSF mechanisms.

Each branch in late fusion process data from sensors independently and does not depend on specific network architecture. Compared with other fusion mechanisms, late fusion is highly flexible. For instance, late fusion can easily combine image-based object detectors and LiDAR-based ones. Late fusion does not involve hidden feature interaction, which also leads to higher efficiency.

Deep fusion involves frequent interactions among hidden features from different branches to gain rich semantic information. Suppose that the depth of branch i is greater than that of branch j, when only one feature fusion is performed, the deep fusion can be formulated as:

$$\begin{aligned} f^L = F^{(L)}(\cdots(F^{(L_i^*+1)}(F^{(L_i^*)}(\cdots(F^{(0)}(M_i))) \\ \oplus F^{(L_j^*)}(\cdots(F^{(0)}(M_j)))))), \end{aligned} \tag{3}$$

where L_i^*, L_j^* denotes that the fusion starts from ith and jth hidden layers, respectively.

Weak fusion does not fuse the hidden features nor fuse the output results. Instead, weak fusion adopts rule-based methods to transform data from one branch to guide the process of data in another branch. The process of weak fusion can be described as:

$$f^L = F^{(L)}(F^{(L-1)}(\cdots F^{(0)}(G(M_i) \oplus M_j))), \tag{4}$$

where G is the function that extracts the guidance from branch i. One typical example of weak fusion is extracting the frustums in the point cloud data using the 2D detection bounding boxes from the image as guidance [35, 44].

Table 1: The collected MSF systems. Performance of each system is evaluated by task-specific metrics (detailed in Sec. 3.3).

System	Task	Fusion	Year	Modality	Performance
EPNet [21]	Object Detection	Deep	2020	C+L	82.70
FConv [44]	Object Detection	Weak	2019	C+L	79.06
CLOCs [33]	Object Detection	Late	2020	C+L	79.70
JMODT [19]	Object Tracking	Deep	2021	C+L	86.12
DFMOT [43]	Object Tracking	Late	2022	C+L	80.17
TWISE [22]	Depth Completion	Deep	2021	C+L	1009.64
MDANet [25]	Depth Completion	Deep	2021	C+L	898.39

3 BENCHMARK CONSTRUCTION

3.1 Benchmark Collection

To collect as many appropriate AI-enabled MSF perception systems as possible for our study, we mainly focus on two sources: (1) the leaderboard of KITTI benchmark [16], and (2) existing MSF-related literature. KITTI is a public autonomous driving benchmark that involves several different perception tasks. For MSF-related literature, we collect papers published in relevant top-tier conferences and journals during the last four years, covering software engineering, robotics, computer vision, etc. We refer readers to our supplementary website [15] for a complete list of selected venues. Eventually, we selected 7 state-of-the-art MSF systems from these two sources based on the following criteria:

- **Multi-sensors**. An MSF system should involve two or more types of different sensors.
- **Open-source**. An MSF system should be open-source so that we can conduct experimental evaluations and enable further replication studies.
- **Data available**. An MSF system should have open-source data for training and evaluation.
- **Representative task**. An MSF system should be designed for representative perception tasks with real-world applications, e.g., object detection.

Table 1 summarizes the seven MSF systems selected in our benchmark. These seven systems cover three different tasks and three different fusion mechanisms. Due to the page limit, we refer audiences to our supplementary website [15] for details of each MSF system.

3.2 Corruption Patterns

Operational environments of many MSF systems are usually open with unexpected condition changes compared with environments during the design phase. Such environment changes are more critical to AI-enabled MSF systems due to the data-driven nature of ML and DL. For instance, an autonomous driving system's object detector might be trained with data only collected from sunny days. While the autonomous driving system is expected to be safe and reliable during rainy days, however, it is hard to determine to what extent the system can handle such a weather change. That is, the weather change in the open environment could result in corrupted sensor signals, leading to potential distribution changes of data that affect an MSF system's performance. To evaluate an MSF system's performance against such operational environments' changes, collecting and labeling real-world data is ideal but not feasible. To address these, we collect and design thirteen corruption patterns (Table 2) to synthesize corrupted signals for MSF systems, which can be grouped into three categories: (1) weather corruption, (2) sensor corruption, and (3) sensor misalignment. Weather corruptions represent the external environment changes of an MSF system, e.g., rainy/foggy days and bright/dark light conditions for a self-driving car, a UAV, etc. Sensor corruptions reflect the internal environment changes of an MSF system, such as transmission noise. Sensor misalignment is specifically designed for MSF systems given that the fusion of different signals requires accurate temporal and spatial calibration. Now we briefly introduce each category.

Table 2: Corruption patterns used in this study.

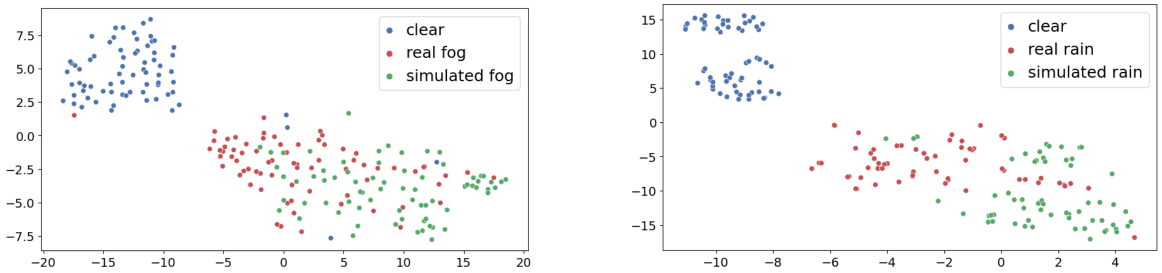
Category	Corruption	Modality
Weather Corruption	Rain (RN)	Camera & LiDAR
	Fog (FG)	Camera & LiDAR
	Brightness (BR)	Camera
	Darkness (DK)	Camera
Sensor Corruption	Distortion (DT)	Camera
	Motion Blur (MB)	Camera
	Defocus Blur (DB)	Camera
	Image Gaussian Noise (GN)	Camera
	Point Cloud Gaussian Noise (GN)	LiDAR
	Image Impulse Noise (IN)	Camera
	Point Cloud Impulse Noise (IN)	LiDAR
Sensor Misalignment	Spatial Misalignment (SM)	Camera & LiDAR
	Temporal Misalignment (TM)	Camera & LiDAR

3.2.1 Weather Corruption. Weather conditions are an important factor that can inevitably affect the sensor's perception in the open environment, resulting in the performance degradation of MSF systems. For example, normal cameras could hardly perceive the surroundings at night. In this work, we leverage weather corruption patterns from two perspectives: (1) light conditions change, and (2) adverse weather conditions.

Lighting conditions. The camera is sensitive to lighting conditions, variations in daylight and road illumination can easily affect the image quality, while lighting conditions' effects on LiDAR are limited [12]. Therefore, we mainly focus on adjusting the **brightness (BR)** and **darkness (DK)** of the image pixels.

Weather conditions. Adverse weather can cause asymmetric measurement distortion of sensors, which poses a significant challenge for MSF perception systems that rely on redundant information. For example, on rainy days, raindrops could lead to pixel attenuation and rain streaks on the image, meanwhile, the droplets will make the laser scattering and absorption, resulting in a lower intensity of points and perceived quality of LiDAR.

In our benchmark, we choose the domain-specific physical model to simulate the properties of two representative adverse weather, i.e., **rain (RN)** and **fog (FG)**. Specifically, we adopt rain model described in [17] and fog model described in [23] for camera, and rain/fog model described in [26] for LiDAR. Another critical problem when designing rain or fog corruptions is ensuring different sensors are sensing identical environments, e.g., the camera and LiDAR are



(a) Fog

(b) Rain

Figure 3: Feature visualization of simulated rain/fog by T-SNEs.

both sensing a rain of 10mm/h. To address this, we control the environmental parameters in LiDAR and camera model to ensure the consistency of the rain's volumes or fog's maximum visibility.

Realisticness validation of Rain and Fog corruption. To validate the naturalness of rain and fog corruptions, we train deep fusion-based classifiers to distinguish real rain/fog scenes from clean scenes using datasets collected from real rainy/foggy weather [3, 4]. Then, we use these trained classifiers to make predictions on simulated data to measure the similarity between simulated and real data. We retrain each classifier five times and take the averaged accuracy. The average classification accuracy of these trained weather classifiers is 98.6% and 98.0% on simulated fog/rain data. These results confirm that the simulated fog/rain data are highly similar compared to the real data.

We further analyze the similarity between the semantic features' distribution of real and simulated data. Specifically, we extract the high-level semantic features from the trained classifier. Then, we utilize T-SNE [41] to reduce the dimensionality of acquired features to 2 and visualize these 2D features. As shown in Figure 3, the distributions of the real and simulated corruptions are similar.

3.2.2 Sensor Corruption. Sensor corruptions reflect internal environment changes that lead to corrupted sensor signals, e.g., noises during transmission, and sensor artifacts that lead to blurry images. In this benchmark, we consider sensor corruption from two perspectives: (1) noise pattern, and (2) sensor artifacts.

Noise Pattern. Noise typically exists in both camera and LiDAR [11]. There are two main sources of noise, one is from the sensor itself, such as sensor vibration [42], random reflections and the low-ranging accuracy of LiDAR lasers [28]. The other is due to the digital signal in its transmission recording process [45]. We leverage two of the most common noise for each sensor, i.e., **Gaussian noise (GN)** and **impulse noise (IN)**. Specifically, Gaussian noise applies Gaussian distributed noise to each point's coordinate in a point cloud or each pixel's value in an image. Impulse noise applies deterministic perturbations to a subset of points or randomly changes the value of image pixels.

Sensor Artifacts. Sensor corruption could also result in artifacts of sensing results. For instance, **defocus blur (DB)** occurs when a camera is out of focus [18]; **motion blur (MB)** appears when a camera is shaking or moving quickly [18]. **Distortion (DT)** is one of the common basic optical aberrations caused by the optical design of lenses [48]. Note that, as an early attempt, we only consider artifacts of camera sensors. We leave artifacts of LiDAR sensors, e.g., one of LiDAR's beams is broken, as the future work.

3.2.3 Sensor Misalignment. Well-calibrated and synchronized sensors are a prerequisite for MSF-based perception systems. However, it is not easy to guarantee the perfect alignment of sensors in the real world [12]. Therefore, we design two corruption patterns, **Spatial misalignment (SM)** and **Temporal misalignment (TM)**, to simulate the misalignment between the camera and LiDAR.

Spatial misalignment. MSF system requires an external calibration of each sensor during the assembly process to ensure that the position measured in different coordinate systems can be converted to each other. However, even with well-calibrated sensors, the position of the sensors can inevitably deviate due to mechanical vibrations (e.g., when a self-driving car rides on a bumpy road) and thermal fluctuations [50]. Suppose a 3D point in the LiDAR coordinate is $\mathbf{p}_{li}$ and a corresponding point in the camera coordinate is $\mathbf{p}_{cam}$. The transformation from the LiDAR coordinate to the camera one can be expressed as:

$$\mathbf{p}_{cam} = \mathbf{T}_{\text{velo}}^{\text{cam}} \, \mathbf{p}_{li} \tag{5}$$

where $\mathbf{T}_{\text{velo}}^{\text{cam}}$ is a rigid body transformation matrix. In our experiments, we add a minor rotation (within 2°) to each rotation angle (i.e., roll, yaw, pitch) to simulate spatial misalignment between the camera and LiDAR.

Temporal misalignment. MSF system requires synchronization of sensors to ensure the output from each individual branch is sensed at the same time. In practical scenarios, sensor or transmission failure may cause a delay in one branch, resulting in a temporal misalignment [47]. To simulate temporal misalignment, for a timestamp t_o, we replace the data $M_i(t_o)$ with the $M_i(t_o - \Delta t)$. This could represent a signal delay of Δt second on branch i.

3.3 Evaluation Metrics

Our benchmarks provide specific quantitative performance evaluation metrics for each perception task, including object detection, object tracking, and depth completion. Then, we define robustness evaluation metrics based on these metrics. Below we describe each perception task and the corresponding evaluation metrics.

Object detection aims to locate, classify and estimate oriented bounding boxes in the 3D space. Note that in this benchmark, we mainly evaluate the detection of *Car* objects with moderate difficulty. The accuracy of object detection can be measured by IOU (intersection over union) and AP (average precision).

IOU measures the overlap area between a ground-truth 3D bounding box B_g and a predicted 3D bounding box B_p over their union[32]. The computation of IOU can be represented as:

$$IOU = \frac{\text{area}\left(B_p \cap B_g\right)}{\text{area}\left(B_p \cup B_g\right)} \tag{6}$$

In our experiments evaluation, we define a successful detection as an IOU larger than 70%.

AP is used to measure the performance of the overall detection performance, which approximates the shape of the Precision/Recall curve as:

$$\text{AP}|_R = \frac{1}{|R|} \sum_{r \in R} \rho_{\text{interp}}\left(r\right) \tag{7}$$

We apply forty equally spaced recall levels [37], i.e., $R_{40} = \{1/40,\ 2/40,\ \ldots,\ 1\}$. The interpolation function is defined as: $\rho_{\text{interp}}\left(r\right) = \max_{r':r' \geq r} \geq \left(r'\right)$, where $\rho(r)$ gives the precision at r.

Multiple object tracking aims to maintain objects' identities and track their location across data frames over time. The accuracy is measured by MOTA (multiple object tracking accuracy) [2]:

$$\text{MOTA} = 1 - \frac{\sum_t \left(\text{FN}_t + \text{FP}_t + \text{IDSW}_t\right)}{\sum_t \text{GT}_t} \tag{8}$$

where FN_t, FP_t, and IDSW_t are the number of misses, of false positives, and of mismatches, respectively, during a period t. The MOTA can be regarded as a measurement of three different types of errors.

Depth completion aims to up-sample sparse irregular depth to dense regular depth. The depth completion tasks focus on predicting the distance for every pixel in the image from the viewer given LiDAR point cloud and image data. We use the Root Mean Squared Error (RMSE, mm) to measure the distance between the predicted depth and ground-truth value:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^{m} \left(d_p^i - d_g^i\right)^2} \tag{9}$$

where d_p^i, d_g^i are the predicted depth and ground-truth of the ith position, m is the total number of ground-truth.

To further evaluate the robustness of different fusion mechanisms across different MSF systems and tasks, we define the robustness of MSF on a corruption pattern $c \in C$ with severity $s \in S$ as its performance P_c^s relative to P_{clean} (performance on clean data):

$$Rb_c^s = P_c^s / P_{clean} \tag{10}$$

where P is measured by one of the evaluation metrics for the corresponding MSF task, i.e., AP, MOTA, or RMSE [1] and $clean$ represents the clean data. A larger Rb_c means that the system's performance against a specific corruption pattern is closer to the normal performance. Then, we estimate the robustness of an MSF system by averaging over all of the corruption patterns c with severity s, i.e.

$$mRb = \frac{1}{|S|} \sum_{s \in S} \frac{1}{|C|} \sum_{c \in C} Rb_c^s \tag{11}$$

A lower mRb means a higher risk of performance degradation when the MSF system is deployed in the open operational environment.

Note that both Rb_c and mRb can generalize to different MSF systems, tasks, and corruption patterns. In this way, we expect our benchmark and evaluation metrics to be flexible and extensible.

3.4 Dataset

KITTI [16] is one of the most popular autonomous driving datasets, which adopts four high-resolution cameras, a Velodyne HDL-64E LiDAR, and an advanced positioning system to collect data from different real-world driving scenarios. KITTI supports diverse perception tasks, including 3D object detection, 3D object tracking, depth completion, etc. During the paper collection process, we also found that more than two-thirds of the MSF perception systems are evaluated on KITTI. To this end, we use the KITTI as our base dataset to construct KITTI-C to benchmark AI-enabled MSF systems' performance and robustness. Note that, corruption patterns used in this study can also generalize to other datasets, such as Waymo [39] and NuScenes [5].

4 EMPIRICAL STUDY DESIGN

In this section, we introduce our research questions and experimental setup. We first investigate the robustness of existing AI-enabled MSF systems from three perspectives: (1) against corrupted signals (**RQ1**), (2) against spatial/temporal misalignments (**RQ2**), and (3) against partial/complete signal loss (**RQ3**). Then, we investigate the potential of repairing these MSF systems' robustness (**RQ4**).

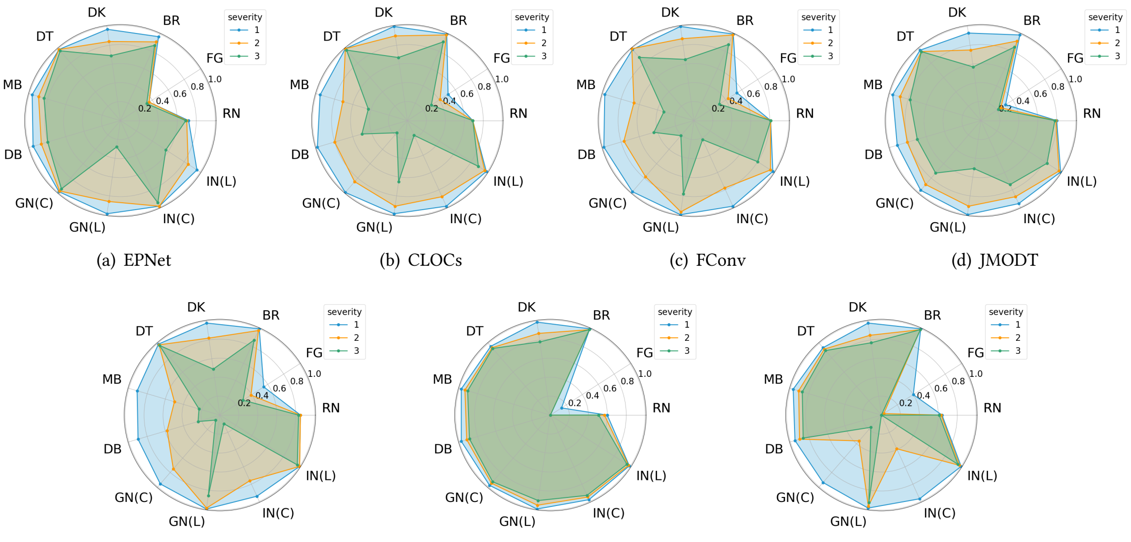
4.1 Research Questions

RQ1. How do AI-enabled MSF-based perception systems perform against common corrupted signals? Though a few AI-enabled MSF perception systems have been proposed and used, there is no systematic study on the robustness of these systems. In this RQ, we focus on corrupted signals due to weather, sensor, and noise corruptions (Table 2). For each corruption pattern, we adopt three different levels of severity. Specifically, for rain and fog, three severity levels represent 10mm/h, 25mm/h, and 50mm/h of rainfall and 104m, 80m, and 51m of visibility, respectively. To sum up, we conduct experiments with 231 different configurations (11 corruptions × 3 levels × 7 MSF systems) to investigate this RQ.

RQ2. How sensitive is AI-enabled MSF when facing spatial and temporal misalignment of sensors? RQ2 aims to evaluate the AI-enabled MSF system's sensitivity to calibration errors. To simulate the spatial misalignment, we rotate the LiDAR sensor around the x, y, and z axes by 0.5°, 1°, and 2°, respectively. To simulate temporal misalignment, we create five levels of LiDAR and camera signal delay, i.e., 0.1s, 0.2s, . . . , 0.5s, respectively. We only investigate temporal misalignment's effects on object tracking systems as the other two tasks are not time-sensitive.

RQ3. To what extent are existing sensing components coupled of an AI-enabled MSF system? This RQ aims to investigate how existing AI-enabled MSF systems are coupled and if they are robust enough against signal loss of one source of signals. To investigate this RQ, we simulate the signal loss with five different levels (10%, 25%, 50%, 75%, 100%) of each branch. For the camera branch, we reshape the image into a one-dimensional array and randomly

[1] Note that we normalize the metric of each task into $[0, 1]$.



(a) EPNet

(b) CLOCs

(c) FConv

(d) JMODT

(e) DFMOT

(f) TWISE

(g) MDANet

Figure 4: Robustness performance of seven MSF systems against different corruption patterns.

Table 3: Average robustness performance of MSF systems against different corruption patterns across three severity levels.

Task		Weather				Sensor			Noise				Rb^{s1}	Rb^{s2}	Rb^{s3}	mRb
		RN	FG	BR	DK	DT	MB	DB	GN(C)	GN(L)	IN(C)	IN(L)				
Object	**EPNet**	0.71	0.35	0.92	0.83	0.98	0.90	0.88	0.98	0.71	0.98	0.79	0.90	0.84	0.72	0.82
	FConv	0.81	0.43	0.96	0.84	0.96	0.66	0.73	0.66	0.92	0.66	0.92	0.93	0.82	0.58	0.78
	CLOCs	0.69	0.41	0.97	0.85	0.99	0.70	0.76	0.67	0.85	0.68	0.95	0.92	0.83	0.58	0.77
Tracking	**JMODT**	0.79	0.26	0.92	0.75	0.97	0.88	0.81	0.86	0.81	0.86	0.94	0.89	0.82	0.70	0.80
	DFMOT	0.84	0.41	0.95	0.77	0.98	0.54	0.57	0.59	0.95	0.60	0.99	0.92	0.78	0.54	0.75
Depth	**TWISE**	0.56	0.05	0.99	0.88	0.94	0.94	0.93	0.95	0.95	0.95	0.98	0.87	0.83	0.79	0.83
	MDANet	0.62	0.14	0.99	0.86	0.92	0.92	0.90	0.49	0.96	0.46	0.98	0.89	0.72	0.64	0.75
Avg		0.72	0.29	0.96	0.83	0.96	0.79	0.80	0.74	0.88	0.74	0.94	0.90	0.81	0.65	0.67

drop pixels. For the LiDAR branch, we randomly remove points with different percentages.

RQ4. What is the weakness of different AI-enabled MSF mechanisms and is it possible to repair them? RQ4 aims to investigate the properties of different fusion mechanisms, and analyze the weakness or potential threats of each based on the experiment results from RQ1-3. We first divide MSF systems into three categories according to their fusion mechanisms. To further investigate the possibility of repairing MSF systems, we make an early attempt on enhancing MSF systems' robustness by improving the fusion mechanism of late and weak fusion.

4.2 Experimental Setup

In experiments, we use Second [46] as the LiDAR branch for CLOCs and DFMOT, Cascade-RCNN [6] as the camera branch for CLOCs, DFMOT, and FConv. We implement all MSF systems with PyTorch 1.8 and Python 3.7. For each system, we use default configurations to ensure a consistent runtime environment. Table 1 shows the performance of each reproduced system. The detailed settings of each system can be found in supplementary website [15]. All experiments are conducted on a server with an Intel i7-10700K CPU (3.80 GHz), 48 GB RAM, and an NVIDIA RTX 3070 GPU (8 GB VRAM).

5 EXPERIMENTAL RESULTS

5.1 RQ1. AI-enabled MSF is not robust against corrupted signals.

Fig. 4 summarizes the robustness benchmark results for seven AI-enabled MSF perception systems against eleven corruption patterns via radar charts. Each axis in the figure represents the robustness score Rb_c^s against corruption c with severity level s. These results reveal that all the selected AI-enabled MSF systems have robustness issues against corrupted signals, while their robustness properties could be varied. For instance, all the selected systems perform poorly against fog (FG) corruption. However, for the blur effects (MB, DB), some systems perform relatively robust (e.g., EPNet, TWISE, JMODT), while some face severe robustness issues (e.g.,

Table 4: Robustness performance of MSF systems against spatial misalignment.

Axis		EPNet	FConv	CLOCs	JMODT	DFMOT	TWISE	MDANet
X	0.5°	0.93	0.92	0.96	0.98	0.99	0.94	0.94
	1°	0.80	0.79	0.83	0.94	0.99	0.82	0.83
	2°	0.46	0.48	0.57	0.80	0.72	0.57	0.58
Y	0.5°	0.45	0.41	0.41	0.93	0.84	0.73	0.77
	1°	0.09	0.04	0.04	0.54	0.75	0.16	0.34
	2°	0	0	0.06	0	0	0	0
Z	0.5°	0.93	0.92	0.96	0.98	0.99	0.94	0.94
	1°	0.80	0.79	0.83	0.95	0.99	0.82	0.83
	2°	0.44	0.48	0.56	0.80	0.72	0.57	0.59
Avg		0.54	0.54	0.58	0.77	0.77	0.62	0.65

CLOCs, FConv, DFMOT). To further analyze how different MSF systems perform against different categories of corrupted signals, we interpret the detailed robustness performance in Table 3 by presenting the average performance against each corruption pattern across three severity levels.

Weather Corruption. As shown in Table 3, weather corruptions pose significant robustness issues for MSF systems, where the average robustness score against rain (RN) and fog (FG) are 0.72 and 0.29, respectively. We also find that the depth completion systems (i.e., TWISE, MDANet) hardly work on foggy days. Specifically, the highest robustness score among depth completion systems is only 0.14. Besides, decreasing brightness affects MSF systems more significantly compared with increasing brightness, where the average robustness scores are 0.83 and 0.96, respectively.

Sensor Artifact. While all the MSF systems are relatively robust against distortion (robustness score higher than 0.9), some of them (i.e., FConv, CLOCs, DFMOT) particularly have significant performance degradation against blur effects (MB, DB). We further qualitatively check the image signals corrupted by distortion (DT) and find that only the edges of images are distorted. This could be one possible reason that the effects of DT are limited.

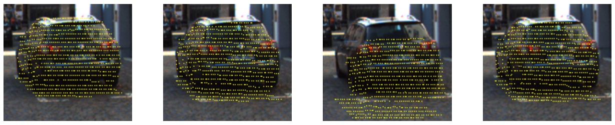
Noise Corruption. As shown in Table 3, camera signals corrupted by noise patterns are usually more vulnerable in MSF systems, where the robustness score against noises in cameras (74.4 (GN), 74.2 (IN)) are lower than those in LiDAR (87.9 (GN), 89.6 (IN)). Based on these observations, adding appropriate filters for image signals could be important for designing robust AI-enabled MSF.

Answer to RQ1: Existing AI-enabled MSF systems are not robust enough against common corruption patterns. Moreover, among the 11 common corruptions, adverse weather causes the most severe robustness degradation.

5.2 RQ2. AI-enabled MSF is sensitive to sensor misalignment.

Through our investigation of RQ2, we find that AI-enabled MSF systems are sensitive to both spatial and temporal misalignment.

Spatial misalignment. Table 4 shows the experimental results of spatial misalignment, where each cell represents the robustness score. According to the average robustness score across different rotation axes and angles (last row of Table 4), we can find that spatial misalignment significantly affects MSF's robustness. Specifically, the highest average robustness score among the seven systems is lower than 0.78. We also find that MSF systems of different tasks could have different sensitivity regarding spatial misalignment. For instance, the robustness scores of object detection systems (i.e., EPNet, FConv, CLOCs) are relatively lower than object tracking and depth completion systems.



(a) Clean (b) X-axis (c) Y-axis (d) Z-axis

Figure 5: An example of a 2° rotation error in calibration around X-, Y-, and Z-axes.

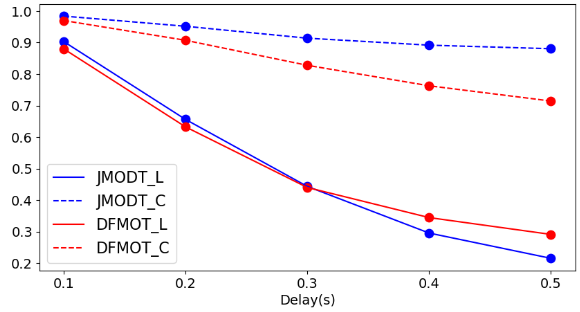
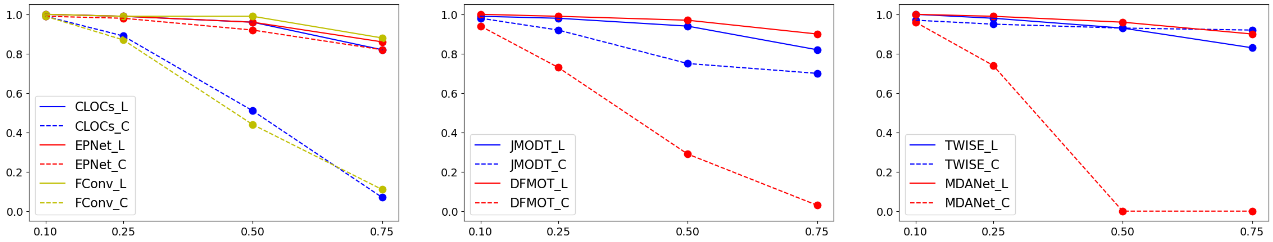


Figure 6: Robustness performance of MSF systems against temporal misalignment.

In addition, we also find that MSF systems are more sensitive to rotation around Y-axis. When the rotation angle around Y-axis is increased to 2° (highlighted in Table 4), five out of seven systems crash (robustness score is 0), while the other two also have poor performance. By contrast, there is no such dramatic decrease for rotations around X- and Z-axes. We qualitatively compare the effects of 2° rotation around different axes by projecting the point cloud onto the image in Fig. 5. A 2° rotation around Y-axis results in a significant malposition between the image and point cloud, which possibly leads to the system crash.

Temporal misalignment. Fig. 6 shows the effects of temporal misalignment on AI-enabled MSF systems for object tracking (i.e., JMODT, DFMOT). As we can observe from Fig. 6, both the camera and LiDAR branch are sensitive to the delay. When the delay increases, the robustness score of the MSF system decreases. In particular, we find that LiDAR is more sensitive (solid lines in Fig. 6) to the delay. When the delay of LiDAR increases to 0.3 seconds, the robustness score of JMODT and DFMOT drops nearly 60% (from 1.0 to 0.4). In contrast, the same level delay of the camera only drops their robustness performance by 10%∼20%.

Answer to RQ2: AI-enabled MSF perception systems are sensitive to both temporal and spatial misalignment, especially for LiDAR. Even small synchronization (0.3 seconds) and calibration errors (2°) can lead to a crash of AI-enabled MSF systems.



(a) Object Detection (b) Object Tracking (c) Depth Completion

Figure 7: MSF systems' performance when partially losing one source of the signals.

5.3 RQ3. Tightly-coupled AI-enabled MSF could be less robust.

When deploying an AI-enabled MSF system, developers might expect it to be reliable even if one of the signals is lost. However, our experiments demonstrate that AI-enabled MSF systems are less robust as they crash when they partially or completely lose a source of signals. Fig. 7 shows the robustness of different MSF systems with different severity levels of signal loss. These results suggest that partially losing either camera or LiDAR signal could affect the MSF system's performance, while losing the camera signal could be more critical (dashed-line in Fig. 7). Specifically, we find that losing the camera signal significantly affects 6 out of 7 systems (except EPNet) compared with losing the LiDAR signal. When losing 75% of the camera signal, 4 out of 7 selected systems have a low robustness performance (*mRb* smaller than 0.2). These results also suggest that existing MSF systems heavily depend on camera signals.

To further investigate AI-enabled MSF systems' robustness against signal loss, Table 5 shows the robustness performance of these systems when completely losing one source of the signal. We can find that when losing LiDAR signals, all of the systems crash. When losing camera signals, 3 out of 7 systems also crash and 2 systems have poor performance (e.g., EPNet, JMODT). Surprisingly, we find that MDANet does not crash when completely losing the camera signal, however, it crashes when losing partial signals (see Fig. 7c). One possible explanation is that due to the sparsity of objects in the image data, discarding 50% or 75% pixels could have dropped all the valuable information (e.g., pixels including objects). The remaining pixels, instead, could bring interference to the MSF system and thus lead to the system crash.

Answer to RQ3: AI-enabled MSF systems could be vulnerable when partially or completely losing one source of signals, even if the other source is working properly. In particular, partially losing camera signals could be more critical for AI-enabled MSF systems. We also find that though tightly-coupled AI-enabled MSF systems have promising performance, they could be less robust when completely losing either camera or LiDAR signals.

Table 5: MSF systems' performance when completely losing one source of the signals.

Modality	EPNet	FConv	CLOCs	JMODT	DFMOT	TWISE	MDANet
C	0.23	0	0	0.13	0	0.50	0.58
L	0	0	0	0	0	0	0

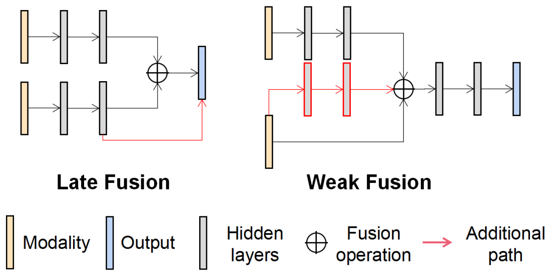


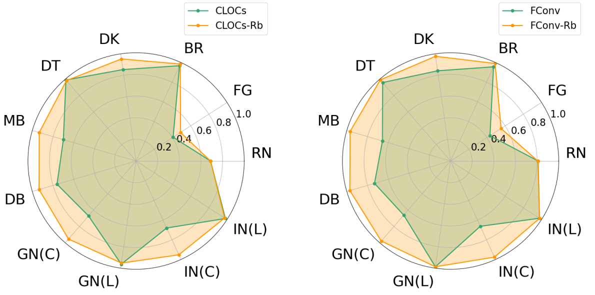
Figure 8: Improved AI-enabled MSF mechanisms.

5.4 RQ4. Fusion mechanisms could affect AI-enabled MSF's robustness and reliability.

While there is no systematic evidence indicating that one specific fusion mechanism is the most robust and reliable, we particularly find that different fusion mechanisms may have unique advantages and potential threats due to their inherent properties. According to our findings from RQ1, three deep fusion MSF systems (i.e., EPNet, JMODT, TWISE) are more robust against blur images (MB, DB) and noise patterns (IN(C), IN(L)) than others (Table 3). According to our finding from RQ3, these systems also perform robustly when partially losing camera signals (Fig. 7). Two late fusion MSF systems (i.e., ClOCs, DFMOT) show similar trends against corrupted signal (RQ1) and signal loss (RQ3). To further investigate the effect of the fusion mechanism on the robustness, we try to repair the badly performed late- and weak-fusion MSF system based on the inherent properties of different fusion mechanisms.

To improve the late fusion, we leverage a shortcut between the LiDAR branch and the fusion layer to enhance the MSF robustness (left part of Fig. 8). Specifically, we design a matching method to aggregate high confidence and unique results from an individual branch to the fusion results. This is motivated by our findings in RQ1 and RQ3, where the camera is more susceptible to external environmental interference.

Weak fusion uses a cascade architecture to connect two modules in series. Its robustness performance bottleneck is due to inaccurate/missing guidance signals. Therefore, for weak fusion, we leverage a neural network to extract extra guidance from another



(a) CLOCs (b) FConv

Figure 9: Performance of the original and enhanced MSF.

Table 6: Improved performance of CLOCs-Rb and FConv-Rb against partial or complete signal loss.

Systems	Modality	10%	25%	50%	75%	100%	Avg
CLOCs-Rb	C	-0.01	0.07	0.41	0.86	0.94	0.45
	L	-0.01	-0.01	0.00	-0.01	0	0.00
FConv-Rb	C	0.00	0.10	0.52	0.86	0.99	0.49
	L	0.00	0.00	0.00	0.00	0	0.00

modality and connect it to the downstream module as an additional guidance branch (right part of Fig. 8). Specifically, we first train a 2D detector by projecting the point cloud to 2D front view images. Then we use the detecting results from the 2D front view as an extra guidance input.

To evaluate the effectiveness of improved fusion mechanisms, we choose CLOCs and FConv as late and weak fusion systems and conduct the same experiments in RQ1 and RQ3. Fig. 9 shows the performance against corruptions of original MSF and enhanced MSF. We find that the enhanced MSF systems are significantly more robust against common corruption patterns. Furthermore, Table 6 shows the improved performance ($\tilde{Rb} - Rb$, where $\tilde{Rb}$ and Rb are robustness score with/without improved fusion mechanisms, respectively) against signal loss. We find that enhanced CLOCs (CLOCs-Rb) and FConv (FConv-Rb) show promising robustness performance against partial and even complete image signal loss. For instance, when the camera signal is completely lost (100% in Table 6), the proposed robustness enhancement strategy almost fully recovers the MSF systems' performance (highlighted in red in Table 6).

Answer to RQ4: MSF systems with the same type of fusion mechanisms may have similar robustness issues due to their inherent properties. Deep fusion performs better against some of the corruption patterns. However, weak fusion and late fusion are easier to be repaired when facing specific robustness issues.

6 DISCUSSION

Discussions. According to our findings from RQ1-3, existing AI-enabled MSF systems are not robust enough. First, corrupted signals could result in significant performance degradation of AI-enabled MSF systems. The data-driven nature makes it challenging to train a robust MSF system that satisfies safety and reliability requirements under all conditions. Therefore, more research on the continuous enhancement of AI-enabled MSF is needed, such as debugging and repair. Our findings from RQ2 also reveal that AI-enabled MSF systems are sensitive to calibration and synchronization errors. In the real world, these two types of errors commonly exist. Even well-calibrated sensors can still be misaligned due to the changes in external environments. To deploy a reliable AI-enabled MSF system, developers must address the calibration issues carefully.

Modular redundancy is a critical way to improve system quality and reliability [13, 14]. By coupling multiple sensors, AI-enabled MSF systems are expected to be robust against signal loss from one specific sensor. However, our experimental results suggest that existing work usually ignores taking this into account when designing AI-enabled MSF, resulting in a lack of robustness. Thus, future work should consider designing AI-enabled MSF systems that can still be reliable with one or more sources of signal loss.

Though existing AI-enabled MSF systems are not robust enough, we also find it possible to repair them with fusion mechanisms improvements. In Sec. 5.4, we propose a potential repairing strategy to repair weak and late fusion mechanisms. The experimental results demonstrate their effectiveness, showing that improving fusion mechanisms could be a promising research direction.

Future Directions. Based on these insights, we summarize the following future directions:

- In this work, we focus on AI-enabled MSF perception systems. However, MSF can also be used in systems beyond perception and autonomous driving. Therefore, more comprehensive benchmarks and more fine-grained robustness evaluation metrics for AI-enabled MSF systems can be considered in the future.
- There is an urgent need for robustness enhancement techniques to continuously improve the reliability of AI-enabled MSF systems. Based on our investigation results, improving fusion mechanisms to repair MSF systems could be a promising research direction.
- Different fusion mechanism-based MSF systems show different robustness issues. Therefore, practical software and system engineering approaches (e.g., testing, debugging, formal analysis, and repairing) would be needed for different MSF systems.

Threats to Validity. In terms of *construct validity*, ideally, it would be highly desirable to expose to diverse and as many corruption datasets as possible, to better approximate the robustness performance of MSF systems. Besides, randomness could also affect the process of synthesizing corrupted data. Therefore, we try our best and adopt a large-scale systematic corrupted dataset (across thirteen corruption patterns and multiple severity levels) to comprehensively measure and analyze the robustness and reliability of MSF in our benchmark. Even though, the robustness results might still not generalize to cases of more diverse types of corruption patterns that are not evaluated in this paper. In terms of *internal validity*, one potential threat is that the leveraged weather corruption may differ from real-world weather. To mitigate this threat, we choose the domain-specific physical model to simulate the properties of adverse weather for different sensors. Further, we ensure that different sensors are sensing identical environments by controlling the hyperparameters in the physical model. In terms of *external validity*, one potential threat is that our analysis results may not be generalized to other MSF systems. To mitigate this threat, we try our best to collect a diverse set of MSF systems with different perception tasks, model structures, and fusion mechanisms.

7 RELATED WORKS

Multi-sensor Fusion. A pioneering work of AI-enabled MSF is MV3D [8]. MV3D takes multi-view representations (i.e., front-view and bird's eye view) of 3D point clouds and images as input and uses a deep fusion mechanism to combine region-wise features from multiple views. To avoid information loss in generating view through perspective projections, EPNet [21] proposes a LiDAR-guided Image Fusion (LI-Fusion) module that enables the interaction between the hidden features of the point cloud and image data to improve system performance. CLOCs [33] is another representative work of late fusion, which leverages geometric and semantic consistencies of 2D and 3D output candidates to produce more accurate final detection results. One of the early works of weak fusion is F-PointNets [35], which uses 2D bounding boxes as guidance to extract frustum in the point cloud and then estimate 3D bounding boxes. FConv [44] extends the F-PointNets by proposing a sliding frustums method to aggregate local point features into frustum-level feature vectors to achieve end-to-end prediction. However, few benchmarks are available to measure the robustness and reliability of these well-designed MSF systems in open environments with corrupted/misaligned sensor signals.

Robustness Benchmarks. Several specific robustness benchmarks designed for one data modality have been proposed. ImageNet-C [18] evaluates the robustness of image specific recognition models against several corruptions. Cityscapes-C [30] extends this ImageNet-C to 2D object detection. However, the weather corruption in ImageNet-C and Cityscapes-C is not guaranteed to respect the underlying physics of weather conditions. Moreover, Mirza et al. [31] evaluate the performance of autonomous driving systems under image data collected in real weather conditions. However, they do not provide a benchmark of LiDAR-based sensing modules against adverse weather conditions. Inspired by ImageNet-C, ModelNet40-C [38] measures the performance of 3D point cloud recognition models. However, these corruptions can only be applied to object-level point clouds instead of open scenes. None of these existing works has focused on benchmarking MSF systems with corrupted data from multiple different modalities. Our benchmark is thus proposed to address this.

MSF Testing and Attack. Zhong et al. [49] propose an evolutionary-based search framework to detect fusion errors for advanced driver assistance systems. Our work is parallel to them, which is to establish a general benchmark rather than testing a specific system. In addition, some recent work has investigated how to attack AI-enabled MSF systems [1, 7, 27, 40]. Cao et al. [7] and Tu et al. [40] attack all branches of MSF systems by inserting adversarial objects. Abdelfattah et al. [1] and Liu et al. [27] investigate attacks on weak fusion and deep fusion systems, respectively. In contrast, our benchmark aims to evaluate the robustness of the MSF systems against common real-world corruptions instead of artificial adversarial objects or perturbations.

8 CONCLUSION

In this paper, we present an early public robustness benchmark of AI-enabled MSF systems, which can further be used as a fundamental evaluation and testing framework for understanding MSF systems' limitations and potential risks. We further perform large-scale robustness evaluation on seven MSF systems against different corruption patterns including *corrupted signals*, *sensor misalignment*, and *signal loss*. Our findings reveal that existing AI-enabled MSF are usually tightly-coupled and not robust enough. Thus, we make an early attempt to enhance the MSF system's robustness by improving fusion mechanisms. Finally, we present discussions and highlight several possible future directions in order to build robust and reliable MSF systems with the emergence of AI.

DATA AVAILABILITY

Our benchmark, replication packages, and detailed evaluation results are publicly available at https://sites.google.com/view/ai-msf-benchmark .

ACKNOWLEDGEMENT

We would like to thank anonymous reviewers for their constructive comments. This project was partially funded by the National Key R&D Program of China (2021ZD0112903) and the National Natural Science Foundation of China under Grant No. 62002158.

REFERENCES

[1] Mazen Abdelfattah, Kaiwen Yuan, Z Jane Wang, and Rabab Ward. 2021. Towards universal physical attacks on cascaded camera-lidar 3d object detection models. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3592–3596.

[2] Keni Bernardin and Rainer Stiefelhagen. 2008. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing* 2008 (2008), 1–10.

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11682–11692.

[4] Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Zou Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. 2022. Boreas: A Multi-Season Autonomous Driving Dataset. *arXiv preprint arXiv:2203.10168* (2022).

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[6] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6154–6162.

[7] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 176–194.

[8] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1907–1915.

[9] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. 2021. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems* 23, 2 (2021), 722–739.

[10] Alex Davies. 2019. Tesla's Latest Autopilot Death Looks Just Like a Prior Crash. https://www.wired.com/story/teslas-latest-autopilot-death-looks-like-prior-crash/.

[11] Yao Duan, Chuanchuan Yang, Hao Chen, Weizhen Yan, and Hongbin Li. 2021. Low-complexity point cloud denoising for LiDAR by PCA-based dimension reduction. *Optics Communications* 482 (2021), 126567.

[12] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems* 22, 3 (2020), 1341–1360.

[13] International Organization for Standardization. 2011. ISO 26262:Road vehicles - Functional safety.

[14] International Organization for Standardization. 2019. ISO/PAS 21448: Road vehicles - Safety of the intended functionality.

[15] Xinyu Gao, Zhijie Wang, Yang Feng, Lei Ma, Zhenyu Chen, and Baowen Xu. 2023. https://sites.google.com/view/ai-msf-benchmark.

[16] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

[17] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. 2019. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10203–10212.

[18] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).

[19] Kemiao Huang and Qi Hao. 2021. Joint Multi-Object Detection and Tracking with Camera-LiDAR Fusion for Autonomous Driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6983–6989.

[20] Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. 2022. Multi-modal Sensor Fusion for Auto Driving Perception: A Survey. *arXiv preprint arXiv:2202.02703* (2022).

[21] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. 2020. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision*. Springer, 35–52.

[22] Saif Imran, Xiaoming Liu, and Daniel Morris. 2021. Depth completion with twin surface extrapolation at occlusion boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2583–2592.

[23] Hans Israël and Fritz Kasten. 1959. Koschmieders theorie der horizontalen sichtweite. In *Die Sichtweite im Nebel und die Möglichkeiten ihrer künstlichen Beeinflussung*. Springer, 7–10.

[24] Shinpei Kato, Shota Tokunaga, Yuya Maruyama, Seiya Maeda, Manato Hirabayashi, Yuki Kitsukawa, Abraham Monrroy, Tomohito Ando, Yusuke Fujii, and Takuya Azumi. 2018. Autoware on board: Enabling autonomous vehicles with embedded systems. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 287–296.

[25] Yanjie Ke, Kun Li, Wei Yang, Zhenbo Xu, Dayang Hao, Liusheng Huang, and Gang Wang. 2021. MDANet: Multi-Modal Deep Aggregation Network for Depth Completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4288–4294.

[26] Velat Kilic, Deepti Hegde, Vishwanath Sindagi, A Brinton Cooper, Mark A Foster, and Vishal M Patel. 2021. Lidar light scattering augmentation (LISA): Physics-based simulation of adverse weather conditions for 3D object detection. *arXiv preprint arXiv:2107.07004* (2021).

[27] Bingyu Liu, Yuhong Guo, Jianan Jiang, Jian Tang, and Weihong Deng. 2021. Multi-view Correlation based Black-box Adversarial Attack for 3D Object Detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1036–1044.

[28] Hongchao Ma and Jianwei Wu. 2012. Analysis of positioning errors caused by platform vibration of airborne LiDAR system. In *2012 8th IEEE International Symposium on Instrumentation and Control Technology (ISICT) Proceedings*. IEEE, 257–261.

[29] Xiaoqian Mao, Wei Li, Chengwei Lei, Jing Jin, Feng Duan, and Sherry Chen. 2019. A brain–robot interaction system by fusing human and machine intelligence. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 3 (2019), 533–542.

[30] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484* (2019).

[31] Muhammad Jehanzeb Mirza, Cornelius Buerkle, Julio Jarquin, Michael Opitz, Fabian Oboril, Kay-Ulrich Scholl, and Horst Bischof. 2021. Robustness of object detectors in degrading weather conditions. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2719–2724.

[32] Rafael Padilla, Sergio L Netto, and Eduardo AB Da Silva. 2020. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*. IEEE, 237–242.

[33] Su Pang, Daniel Morris, and Hayder Radha. 2020. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10386–10393.

[34] Zi Peng, Jinqiu Yang, Tse-Hsun (Peter) Chen, and Lei Ma. 2020. A First Look at the Integration of Machine Learning Models in Complex Autonomous Driving Systems: A Case Study on Apollo *(ESEC/FSE 2020)*. Association for Computing Machinery, New York, NY, USA, 1240–1250. https://doi.org/10.1145/3368089.3417063

[35] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. 2018. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 918–927.

[36] Stamatios Samaras, Eleni Diamantidou, Dimitrios Ataloglou, Nikos Sakellariou, Anastasios Vafeiadis, Vasilis Magoulianitis, Antonios Lalas, Anastasios Dimou, Dimitrios Zarpalas, Konstantinos Votis, et al. 2019. Deep learning on multi sensor data for counter UAV applications—A systematic review. *Sensors* 19, 22 (2019), 4837.

[37] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. 2019. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1991–1999.

[38] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z Morley Mao. 2022. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296* (2022).

[39] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2446–2454.

[40] James Tu, Huichen Li, Xinchen Yan, Mengye Ren, Yun Chen, Ming Liang, Eilyan Bitar, Ersin Yumer, and Raquel Urtasun. 2021. Exploring adversarial robustness of multi-sensor perception systems in self driving. *arXiv preprint arXiv:2101.06784* (2021).

[41] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[42] Rongrong Wang, Bingnan Wang, Maosheng Xiang, Chuang Li, Shuai Wang, and Chong Song. 2021. Simultaneous time-varying vibration and nonlinearity compensation for one-period triangular-FMCW lidar signal. *Remote Sensing* 13, 9 (2021), 1731.

[43] Xiyang Wang, Chunyun Fu, Zhankun Li, Ying Lai, and Jiawei He. 2022. DeepFusionMOT: A 3D Multi-Object Tracking Framework Based on Camera-LiDAR Fusion with Deep Association. *arXiv preprint arXiv:2202.12100* (2022).

[44] Zhixin Wang and Kui Jia. 2019. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1742–1749.

[45] Katja Wolff, Changil Kim, Henning Zimmer, Christopher Schroers, Mario Botsch, Olga Sorkine-Hornung, and Alexander Sorkine-Hornung. 2016. Point cloud noise and outlier removal for image-based 3D reconstruction. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 118–127.

[46] Yan Yan, Yuxing Mao, and Bo Li. 2018. Second: Sparsely embedded convolutional detection. *Sensors* 18, 10 (2018), 3337.

[47] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. 2021. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors* 21, 6 (2021), 2140.

[48] Keyao Zhao, Kang Liao, Chunyu Lin, Meiqin Liu, and Yao Zhao. 2021. Joint distortion rectification and super-resolution for self-driving scene perception. *Neurocomputing* 435 (2021), 176–185.

[49] Ziyuan Zhong, Zhisheng Hu, Shengjian Guo, Xinyang Zhang, Zhenyu Zhong, and Baishakhi Ray. 2022. Detecting multi-sensor fusion errors in advanced driver-assistance systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*. 493–505.

[50] Yufeng Zhu, Chenghui Li, and Yubo Zhang. 2020. Online Camera-LiDAR Calibration with Sensor Semantic Information. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. 4970–4976. https://doi.org/10.1109/ICRA40945.2020.9196627

[51] Chris Ziegler. 2016. A Google self-driving car caused a crash for the first time. https://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report.

Received 2023-02-02; accepted 2023-07-27